Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W645–W648***
DOI: 10.1093/nar/gkh420

BSDD: Biomolecules Segment Display Device—a web-based interactive display tool

P. Selvarani[1], V. Shanthi[1], C. K. Rajesh[1], S. Saravanan[1] and K. Sekar[1,2,*]

[1]Bioinformatics Centre and [2]Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore 560 012, India

Received February 10, 2004; Revised and Accepted April 5, 2004

ABSTRACT

An interactive web-based display tool, Biomolecules Segment Display Device (BSDD), has been developed to search for and visualize a user-defined motif or fragment among the protein structures available in the Protein Data Bank (PDB). In addition, the tool works for the structures available in a selected subset of non-homologous protein structures (25% and 90% sequence identity) . The graphics package RASMOL has been incorporated as an interface to visualize the three-dimensional structure of the user-defined motif. In addition, the software can be used to extract the atomic coordinates of the required fragment and save them to the client system. The atomic coordinates are updated every week from the RCSB–PDB server, and hence the results produced by BSDD are up to date at any given time. The software BSDD is available over the World Wide Web at http://iris.physics.iisc.ernet.in/bsdd or http://144.16.71.2/bsdd.

INTRODUCTION

Owing to the advent of synchrotron radiation facilities and high-power digital computers, there has been an explosive growth in the number of known protein structures. The three-dimensional structures of nearly 25 000 proteins and nucleic acids are currently available in the public-domain Protein Data Bank (PDB) (1). In the post-structural-genomics era, given the wealth of information available in the PDB, predicting three-dimensional structure from an amino acid sequence is one of the crucial problems to be addressed in structural biology (2). The three-dimensional structure dictates the function and the biological role of a protein molecule. It is frequently found that similar amino acid sequence repeats occur in several protein structures, and it is essential to know their location within the entire protein molecule. Sequence repeats within the polypeptide chain might suggest symmetries in structures (3). It is noteworthy that the location of a particular motif in a protein structure may give insights into the possible function (4), and one could fine-tune the three-dimensional structure by means of site-directed mutation experiments to incorporate the desired/required function. In light of the above, there is a need to see the location of a user-defined sequence motif or structural motif within the entire protein/nucleic acid molecule. Towards this end, a web-based software program has been developed to visualize a user-defined sequence motif from the three-dimensional structures available in the PDB and in a set of non-homologous protein structures (both 25% and 90% sequence identity).

METHODOLOGY

The PDB is currently maintained and run by the Research Collaboratory for Structural Bioinformatics (RCSB). The RCSB–PDB (1) file: pdb_seqres.txt (ftp://ftp.rcsb.org/pub/pdb/derived_data/pdb_seqres.txt) contains the single-letter amino acid/nucleic acid sequence for all the proteins and nucleic acids whose three-dimensional structures are available in the PDB. Since the aim of the BSDD software is to show the three-dimensional structure of a user-defined region, the sequence information available in pdb_seqres.txt is inadequate. For example, in the case of PDB id code 1agw (pdb1agw.ent), the number of amino acid residues in the file is 310. However, in the PDB file, the residue numbering starts at 464. In addition, the PDB file does not have atomic coordinates for residues 486–490. In another example, PDB id code 1914 (pdb1914.ent), there are 232 amino acid residues (single polypeptide chain) available in the pdb_seqres.txt file. However, the three-dimensional atomic coordinates in the PDB file have three segments, namely, (i) 2001–2097, (ii) 3001–3008 and (iii) 4004–4081; and the atomic coordinates are not available for the region 2035–2046. In addition, several PDB files contain information about the insertion and deletion of amino acid

*To whom correspondence should be addressed. Tel: +91 080 22933059; Fax: +91 080 23600551; Email: sekar@physics.iisc.ernet.in

This work is dedicated to Professor M. Vijayan, Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

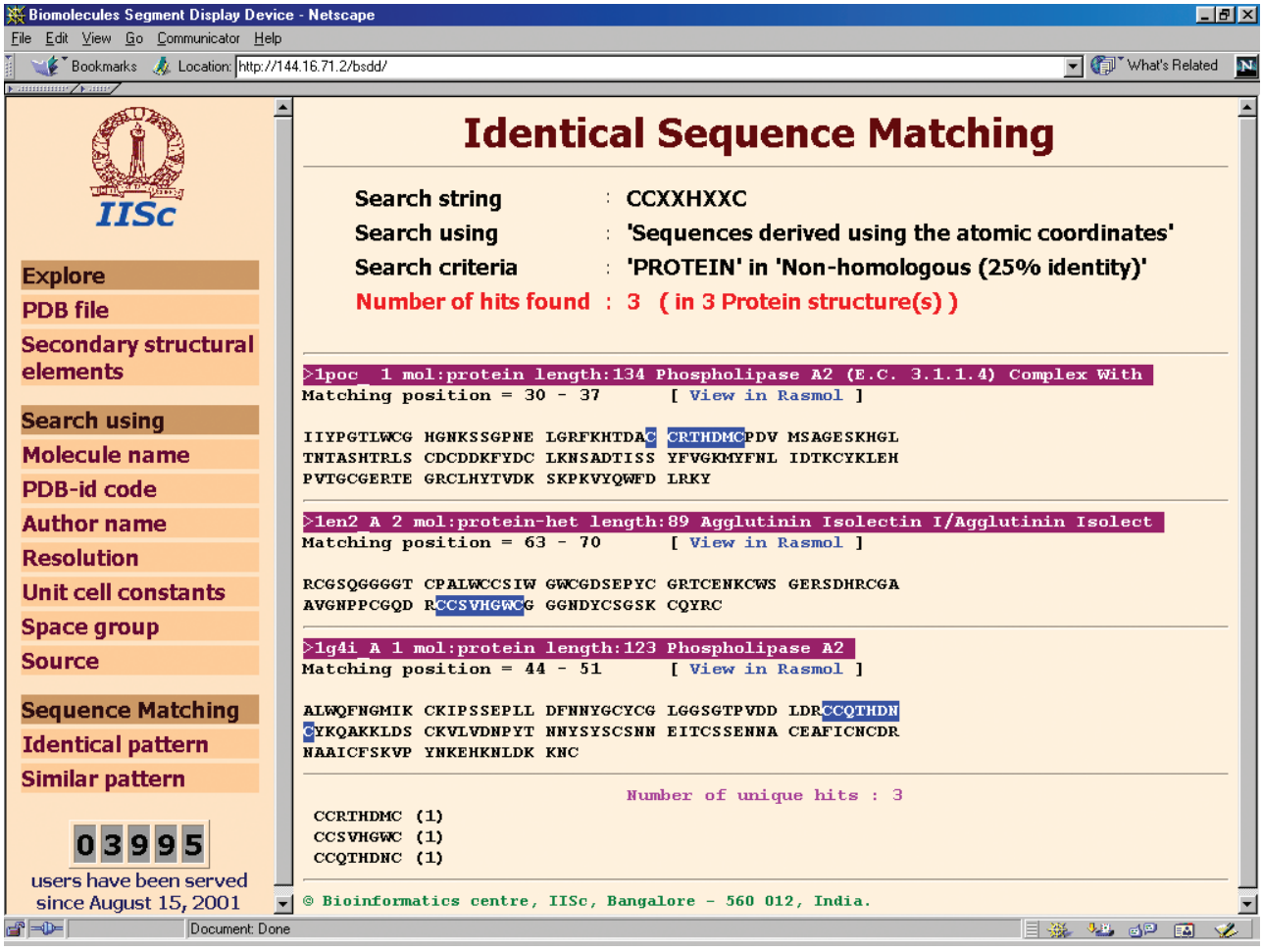

Figure 1. Typical output for pattern matching using 'Sequences derived using the atomic coordinates' from the non-homologous (25% identity) sequence database.

residues. The inserted residues have sequence numbers, e.g. 44A and 44B in the PDB file pbd 104ℓ.ent and 60A and 60B in pdb1a0l.ent. In view of the above, a separate sequence database has been created by comparing the sequence information available in pdb_seqres.txt with the atomic coordinates present in the corresponding protein structure. In the same manner, separate derived databases have also been created for the non-homologous protein structures (25% and 90% sequence identity). The non-homologous dataset was derived by Hobohm and Sander (5).

The search engine is written in the PERL (CGI/PERL) programming language and it runs on our Intel-based Solaris Bioinformatics server (a 3.06 GHz Pentium IV processor, 1 GB of Random Access Memory). The front-end data input part of the tool is written in HTML and JavaScript, which allows user-friendly web forms. The facility is freely accessible without restriction for academic and research purposes over the World Wide Web at http://iris.physics.iisc.ernet.in/bsdd or http://144.16.71.2/bsdd.

UTILITIES AND PERSPECTIVES

The BSDD software offers the following three main utilities for user queries; (i) 'Explore', (ii) 'Search using' and (iii) 'Sequence matching'. The utility 'Explore' has two functions. The functions are 'PDB file' and 'Secondary structural elements''. Using the former function the user can see the header information or the complete content of the PDB file. The same option is also available in the software Structural Classification of Proteins, SCOP (6). However, the BSDD software displays the protein structure that is currently available in the PDB. The function 'Secondary structural elements' culls the user-specified PDB file and displays the information about the secondary structural elements (α-helix and β-sheet) available in the PDB file under the appropriate heading 'HELIX' or 'SHEET'. The output screen contains the amino acid residues involved in each of the secondary structural elements. The user can visualize the three-dimensional structure of the individual secondary structural elements or the location of these within the entire molecule by clicking the button 'View in Rasmol'. The software is optimized for NETSCAPE Version 4.X. To use the RASMOL display, the user needs to install the software only on the first occasion (for instructions, see http://144.16.71.2/bsdd/rasmol.html). In addition, the user can save the three-dimensional atomic coordinates of a particular secondary structural element for further study.

To perform the options 'Explore' and 'Search using', the user needs to supply the four-letter PDB id code of the protein

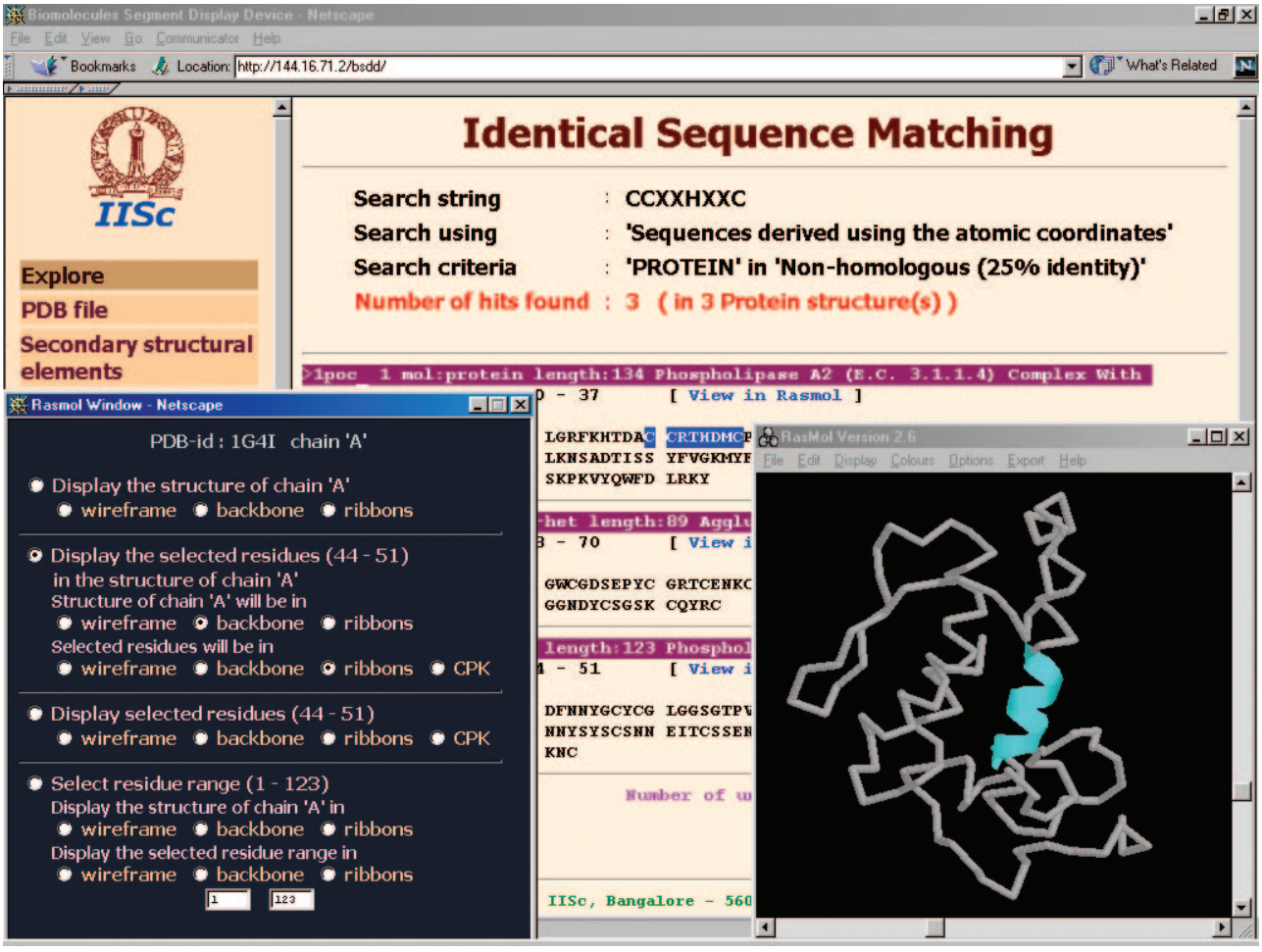

Figure 2. A sample output frame for a consensus pattern search. The string CCXXHXXC is used for a pattern matching search. The search string is shown as a ribbon representation, while only the backbone trace is shown for the remaining part of the protein model. To visualize other regions, the user needs to change the residue number in the menu box provided at the bottom of the RASMOL window box.

structure of interest. The second option, 'Search using', offers seven search facilities, based on (i) molecule name, (ii) PDB id code (four-character PDB identifier), (iii) author name, (iv) resolution, (v) unit cell constants, (vi) space group and (vii) source. These options are self-explanatory, and the user can perform the search without any difficulty. In the first four options, the search can be performed in (i) all the protein structures, (ii) the 25% and (iii) the 90% non-homologous protein structures.

Identical pattern and similar pattern searches are under the 'Sequence matching' option. As pointed out above, two sequence databases are employed for the search facility, namely, (i) sequence information available in pdb_seqres.txt and (ii) sequence information derived from the PDB atomic coordinates file. Users need to choose the latter option if they wish to see the three-dimensional structure of the query motif within the entire protein molecule. Most current search engines perform pattern matching only with the structures available in the PDB, but the BSDD software performs such as search in the non-homologous (25% and 90% sequence identity) protein structures too. At present there are 2063 (25% identity) and 6243 (90% identity) protein chains in the non-homologous category. Our software has several flexible options and can be used to extract the atomic coordinates of the given input string or any part of the protein/nucleic acid structures. The results obtained from the search correspond to the most recent information available in the PDB.

A sample output of a typical sequence pattern search is shown in Figure 1. The search pattern is taken from PROSITE (7) and corresponds to an important consensus pattern involved in the function of phospholipase A_2. The pattern is CCXX HXXC (where C is cysteine, X is any amino acid and H is histidine). As can be seen in Figure 1, there are only three hits that correspond to the consensus pattern of phospholipase A_2. A RASMOL (8) window pops up (Figure 2) on clicking the 'View in Rasmol' link (Figure 1). The RASMOL window has various options for the user to see the region of interest within the entire molecule (Figure 2).

In summary, the software offers several utilities to the user community working in the area of structural biology and bioinformatics. Most importantly, the software allows the user to visualize the location of the query motif within the entire structure or any fragment of a particular motif of the entire molecule (protein/nucleic acid). The non-homologous sequence information (25% and 90%) will be updated as and when updated information is available from Hobohm and Sander's FTP site in Heidelberg, Germany. The atomic

coordinates and the single-letter protein/nucleic acid sequence information are updated every week from the RCSB–PDB, and hence the results produced by BSDD are up to date at any given time. The software will be improved in accordance with the progress of, and requests from, the scientific community around the world.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the facilities at the Supercomputer Education and Research Centre, the Interactive Graphics Based Molecular Modelling (IGBMM) and Distributed Information Centre (DIC). The Department of Biotechnology, Government of India, India supports the facilities IGBMM and DIC.

REFERENCES

1. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
2. Pabo,C.O. (1987) New generation databases for molecular biology. *Nature*, **327**, 467.
3. McLachlan,A.D. and Stewart,M. (1976) The 14-fold periodicity in α-tropomyosin and the interaction with actin. *J. Mol. Biol*., **103**, 271–298.
4. Kasuya,A. and Thornton,J.M. (1999) Three dimensional structure analysis of prosite patterns. *J. Mol. Biol*., **286**, 1673–1691.
5. Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci*., **3**, 522–524.
6. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.
7. Bairoch,A., Bucher,P. and Hofmann,K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res*., **25**, 217–221.
8. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci*., **20**, 374–382.